第 1 章
データの表現・テキスト処理

2023 年 10 月 11 日

学習目標

(1) 仮想デスクトップ (VDI) を利用してみる．
(2) データの表現方法について理解する．
(3) テキストの表現方法について理解する．
(4) 画像・音声・動画の表現方法について理解する．
(5) テキスト処理を体験してみる．

本章は，専修大学商学部の高萩栄一郎の著作である．

1　仮想デスクトップ (VDI) の利用

専修大学では，「VDI（Virtual Desktop Infrastructure）」を利用できます．

1.1　仮想デスクトップ (VDI) とは

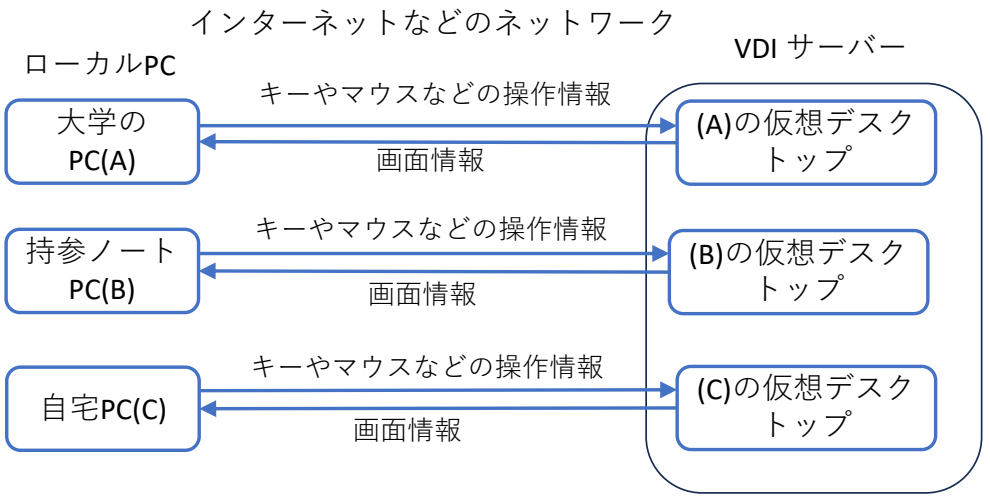

図 1　VDI のイメージ

図 1 は，VDI のイメージです．利用者用の PC(仮想デスクトップ）が VDI サーバーの中に用意されおり，それをみなさんはインターネットを介して利用します．

利用者の PC(ローカル PC) は，インターネットなどを介して，VDI サーバーに接続します．VDI サーバーの中に，接続し

た PC 毎に,「仮想デスクトップ」と呼ばれる環境（1 つの Windows マシンに相当）を作成し，そこで，さまざなアプリケーションを実行します．ローカル PC はその仮想デスクトップにキー入力情報やマウスの操作情報を送り，仮想デスクトップは，さまざな処理をし，その結果の画面情報をローカル PC に送り，ローカル PC は，その画面を表示します．これにより，ローカル PC を利用するのとほぼ同じように仮想デスクトップを利用できます．

専修大学では，マイクロソフト社のクラウドサービス AVD (Azure Virtual Desktop) という商品を利用しています．

ローカル PC との大きな違いはファイルの保存場所です．ローカル PC の場合は，その PC の中 (「ドキュメント」ホルダなど）に保存等ができますが，専修大学の仮想デスクトップでは，基本的に，専修大学 One Drive に保存・読み込みなどをします．そこで，ローカル PC で作業を行うときも，OneDrive にファイルを保存すれば，VDI と同期され，ローカル PC と VDI で保存場所をあまり意識せずにファイルを利用できます．

VDI を利用することにより，大学で利用するソフトウエアをローカル PC にインストールすることなしに利用できたり，利用者全員が同じ環境で PC を利用できたりするなどのメリットがあります．ただし，VDI で OneDrive を利用するとき，VDI にログオンする毎に OneDrive を接続する必要があります (2023/9/10 現在)．

専修大学の VDI では，同時に利用可能な台数に制限があります．そこで，VDI の利用は，たとえば，ローカル PC にインストールされていないソフトウエア利用など，VDI の利用が必要な場合にしましょう．ブラウザの利用や Office ソフトウエアの利用のみなどローカル PC のみの使用で十分な場合は，ローカル PC で利用しましょう．この授業でも本章など一部を除いて，ローカル PC でも学修できるように記述しています．

1.2　仮想デスクトップの利用 (リモートデスクトップアプリの利用)

本授業では，AVD 用のリモートデスクトップアプリをローカル PC にインストールして利用します．その他にブラウザを使って VDI を利用する方法もありますが，リモートデスクトップアプリの利用のほうが安定している印象があるので，こちらを利用します．

アプリのダウンロードとインストールの方法は，仮想デスクトップ（VDI）(https://www.senshu-u.ac.jp/isc/services-list/VDI.html) から WindowsOS 用 AVD マニュアル (PDF ファイル) があります．このマニュアルにしたがって，次の各作業を行ってみてください．

- 1. リモートデスクトップアプリのインストール及び AVD の接続方法
- 2. 各種ブラウザーの初回利用設定方法
- 3. AVD 環境から Office 製品へのサインイン方法
- 4. AVD 環境から OneDrive へのサインイン方法
- 5. AVD 環境・接続元 PC 間のファイル転送方法

1.3　仮想デスクトップの利用 (ブラウザの利用)

ブラウザ (Google Chrome など) から，https://client.wvd.microsoft.com/arm/webclient/index.html で，ログオンします．現在 (2023/09/25 現在)，設定として，新しい顧客 の設定を Off にしないと，日本語入力ができません．新しい顧客 の設定を Off にして，リモートデスクトップに入ったあと，AVD のウィンドウ右上の歯車のマークをクリック，「リモートキーボードレイアウトを選択する」のドロップダウンリストから，「日本語 (101/102 キー)」を選択してください．

1.4　うまく接続等ができない場合

専修大学情報科学センターで VDI 接続の講習会が開かれます (2023 年度は 9・10 月に予定)．講習会に参加して，接続できるようにしておきましょう．
講習会案内の URL(2023 年度):仮想デスクトップ (VDI) 利用説明会のお知らせ

2　データの表現方法

本節では，データ（数値，文字，画像等）の表現方法について学修します．コンピュータのデータは，デジタルデータと呼ばれ，0 と 1 の組み合わせで表現されます．1 桁の 0 または 1 をビット (bit), 通常，8 ビットのまとまりを 1 バイト (byte) といいます．1 ビットでは，0 か 1 の 2 種類の情報を表現でき，8 ビットでは，$2^8 = 256$ 種類の情報を表現できます．

情報入門 2 のホームページにある dexp.xlsx を利用します．オレンジのセルが，値を入力してみるセル，青のセルが皆さんで計算式を入力するセルになっています．

2.1　数値の表現方法

2.1.1　整数型の表現

整数は，2 進数で表現します．10 進数が，$0, 1, \ldots, 8, 9, 10, 11, \ldots$ のように，9 の次，10 になると繰り上がって 10(イチ ゼロ) となる数と同じように，2 進数は，$0, 1, 10, 11, 100, 101, 110, 111, 1000, \ldots$ のように，2 になると繰り上がる数です．

図 2 の (1) 整数型は，B4 に整数を入れると，C4 に 2 進数表現を文字列として表示するものです．Excel では，10 数数 (decimal) の数値を 2 進数 (binary) に変換する関数は，DEC2BIN です．

C4: =DEC2BIN(B4)

B4 に 0 から 10 くらいまで数値を 1 ずつ変化させ，2 進数表現の変化を確認しましょう．

負の数はどのようになるのでしょうか？ B4 に-1 を入れ見ましょう．「1111111111」と 1 が 10 桁表示された思います．これは，整数を 10 ビットで表現したとき，1 を加えると 0 になる数 (ただし，$1111111111 + 1 = 10000000000$ となるが，10 ビットとしたので，最初の 1 は保持されない（無視）) です．このような表現方法を 2 の補数表現と呼ばれます．

	A	B	C	D	E	F
1	(1) 整数型					
2	10進数から2進数					
3		整数	DEC2BIN			
4		9	1001			
5						
6	(2) 指数表現					
7		小数	指数表現			
8		28053	2.80530E+04			
9						
10		仮数	指数	小数		
11		2.8053	2	280.53		
12	※　コンピュータ内部は，10進数ではなく2進数					
13						
14	(3)　計算誤差					
15		整数1	整数2	整数1-整数2	10倍	10を引く
16		53	52	1	10	0
17						
18		小数1	小数2	小数1-小数2	10倍	1を引く
19		5.3	5.2	0.1	1	-3.55271E-15

図 2　数値の表現（dexp.xlsx のシート「数値」）

2.1.2　浮動小数点型の表現

図 2 の (2) 指数表現は，数値を指数表現で表現し直したものです．

- C8: =B8
- C8 を右クリックして，セルの書式設定
- 表示形式のタブで，分類を指数，小数点以下の桁数を 5 と設定します．

B8 に 280.53 と入力すると，「2.80530E+02」と表示されます．「E+02」の右は，10 の何乗倍かを表し，これは，$10^2 = 100$ 倍を表します．「2.80530E+02」は，$2.80530 * 10^2 = 280.530$ を表します．

このような表現をしたとき，2.8053 の部分を仮数，+02 を指数と呼びます．B11 に仮数，C11 に指数を入力したとき，D11 にその数値を表示させましょう．

D11: `=B11*10^C11`

このような表現をすることにより，大きな数（例えば，987654321012.99) や 0 に近い数 (0.0000000000123) などを表現できます（B8 に，大きな数や 0 に近い数を入力すると，表示形式が標準のため，自動的に指数形式で表示されます)．

小数はコンピュータ内部では，2 進数で同様な方法で表現されます．仮数とその符号，指数部の整数の 3 つの部分で表現されます．このような表現方法を浮動小数点型と呼ばれています．

2.1.3 計算誤差

2 進数では正確には，10 進数を表現できません．これは，10 進数で 1/3 を有限の桁の小数で表現できないのと同様の理由です．小数の計算を行うと誤差が生じることがあります．図 2 の (3) 計算誤差で，誤差を生じさせてみます．

整数 1 − 整数 2 = 1 となる数値を B16,C17 に入力し，整数 1 − 整数 2(D16) を 10 倍し (E16),10 を引きます (F16)．整数の場合，誤差なく 0 になります．

小数 1 − 小数 2 = 0.1 となる数値を B19,C19 に入力し，小数 1 − 小数 2(D19) を 10 倍し (E19),10 を引きます (F19)．整数の場合，誤差なく 0 になります．図のように小数 1 に 5.3, 小数 2 に 5.2 を入力したときは，-3.55271E-15 となり 0 に近い

ですが誤差が生じます．（入力した小数により，0 となることもあります）．

このように，小数の計算では誤差が生じることがあるので，会計などの事務計算では注意しましょう．

動画 (復習用教材):計算誤差（音声付）

動画 (復習用教材):計算誤差（音声なし）

2.2 文字の表現

文字には番号が振られており，その番号は文字コードと呼ばれています．英数字小文字は，7 ビットで表現される ASCII コードが使われます．漢字やハングル文字などは，全世界の文字を基本的に 16 ビット (2 バイト) の整数値で表現する ユニコード (Unicode) が使われることが多いです．実際の文字コード (符号化方式) は，UTF-8 と呼ばれる方式で表現されることが普及しています．

文字コードは，4 ビットで 1 桁を表す 16 進数 (hexadecimal) で表示します．ASCII 文字は 8 ビットに納まるので，16 進数 2 桁，日本語文字のユニコードの整数値は，2 バイトなので 16 進数 4 桁で表示します．

文字コード（数値，10 進数）を，16 進数の文字列に変換する関数は DEC2HEX で，16 進数の文字列を文字コードに変換する関数は HEX2DEC です．

図 3 の (1)ASCII コードは，文字コード (32〜126）を入力するとそのコードの文字を表示する関数 char を利用したものと ASCII 文字を入力するとその文字コードを表示する関数 code を利用したものです．

- C3: `=DEC2HEX(B3)`
- D3: `=CHAR(B3)`
- C4: `=CODE(B6)`
- D4: `=DEC2HEX(C6)`

	A	B	C	D	E
1	(1) ASCII コード				
2		文字コード(32～126)	文字コード(16進数)	文字	
3		65	41	A	
4					
5		文字(ASCII)	文字コード(10進数)	文字コード(16進数)	
6		A	65	41	
7					
8	(2) Unicode 文字				
9		日本語文字	Unicode(10進数)	Unicode(16進数)	
10		漢	28450	6F22	
11					
12		Unicode(16進数)	Unicode(10進数)	日本語文字	
13		6F22	28450	漢	
14					
15	(3) UTF-8				
16		日本語文字	URL encode	UTF-8	
17		漢	%E6%BC%A2	E6BCA2	
18					

図 3　数値の表現（dexp.xlsx のシート「文字」）

図 3 の (2)Unicode は，漢字などの文字を入力するとその Unicode の整数値を表示する関数 UNICODE 利用したものと Unicode の整数値を入力すると，その文字を表示する関数 =UNICHAR(C13) を利用したものです.

- C10: `=UNICODE(B10)`
- D10: `=DEC2HEX(C10)`

- C13: =HEX2DEC(B13)
- D13: =UNICHAR(C13)

いろいろな文字コードや文字で試してみましょう.

16進数	0	1	2	3	4	5	6	7
0				0	@	P	`	p
1			!	1	A	Q	a	q
2			"	2	B	R	b	r
3			#	3	C	S	c	s
4			$	4	D	T	d	t
5			%	5	E	U	e	u
6			&	6	F	V	f	v
7			'	7	G	W	g	w
8			(	8	H	X	h	x
9			)	9	I	Y	i	y
A			*	:	J	Z	j	z
B			+	;	K	[	k	{
C			,	<	L	¥	l	|
D			-	=	M	]	m	}
E			.	>	N	^	n	~
F			/	?	O	_	o	

図 4　ASCII コード表 (文字のみ)

図 3 の (3)UTF-8(参考) は，日本語文字を UTF-8 に変換するものです．UTF-8 で日本語文字は，16 進数 3 桁 (3 バイト) に変換されます.

図 4(ワークシート「文字コード表」) は，ASCII コードの文字コード（文字のみ）表です.

表の見方:「M」は，上方 (H1) の 4 が 16 進数の 1 桁目，左方 (B16) の D が 16 進数の 2 桁目になり，「M」は 16 進数で 4D となります.

2.3　画像・音・動画の表現

画像，音声，動画も，数値や文字と同様に，コンピュータ内部ではデジタル表現（2 進数での表現）で格納されています．それぞれどのように，表現されるのか説明します．

2.3.1　画像の表現

画像の表現は，大きく分けると 2 種類あります．

■ラスタ型　ラスタ型は，格子状のマス目を作り，各マス目をピクセル（画素）と呼び，各ピクセルの色番号で表現します．ピクセル数が多ければ，高画質になります．例えば，最近のスマートフォンのカメラのピクセル数は 4800 万ピクセルのものがあります (2023 年 8 月)．拡大していくと，ピクセルが見えるように，拡大の限界があります．また，ファイルサイズが大きくなる傾向にあります．

写真やスキャナーで取り込んだ画像などが当てはまります．ファイル形式では，JPEG, GIF, PNG などがあり，JPEG などでは，ファイルサイズを抑えるため，圧縮が行うことができます．

■ベクター型　ベクター型は，座標を使って，直線や円弧などの曲線や塗りつぶしなどで表現します．
次の囲みの中は，ベクター型の表現方法の 1 つである SVG(Scalable Vector Graphics) での円弧を描く表現です．

これは，「座標 (150,150) を中心に半径 100 の円を太さ 1 の黒線で描き，中を白で塗りつぶせ」という命令の SVG です．

```
<svg width="300" height="300">
  <circle cx="150" cy="150" r="100" stroke="black" stroke-width="1" fill="white" />
</svg>
```

ベクター型は，拡大に強く，拡大しても曲線が保たれます．また，ファイルサイズは比較的小さくなります．しかし，ラスター型のファイルをベクター形式に変換するのは，かなり困難な処理と言われています．

ベクター形式は，線や図形を組み合わせて作成されたイラストや設計図などで使われます．ファイル形式には，SVG, EPS, PDF などがあります．

■比較　maru.bmp は，ラスター形式のファイルで円を描いたものです．2000% くらいまで拡大すると，ピクセルの矩形が見えてきます．

maru_svg.html は，上記 SVG を html ファイルに組み込んだものです．拡大していっても，ピクセルの矩形が見えず，円弧の曲線が保たれているのが確認できます．

動画 (復習用教材):ラスタ型ベクター型比較（音声付）

動画 (復習用教材):ラスタ型ベクター型比較（音声なし）

2.3.2　音の表現

音は，空気の振動の波として表されるので，波の大きさを一定間隔で測定（サンプリング）して，その大きさをデジタル表現します．1 秒間に何回，サンプリングをするのかをサンプリング周波数と呼んでいます．たとえば，44.1kHz（CD のサンプリング周波数）は，1 秒間に 44100 回サンプリングをすることを意味します．

2.3.3　動画の表現

動画は，パラパラ漫画のように画像（フレーム）を短い間隔で切り換えて表示することにより表現しています．1 秒間に何回，フレームを切り換えるのかをフレームレートと呼ばれています．

2.3.4　圧縮

ラスタ型の画像や音，動画は容量が大きくなります．そのままですと，保存や通信に不向きになります．そのとき，圧縮する技術を使って容量を小さくして保存，通信をします．圧縮には，可逆圧縮と非可逆圧縮の 2 種類あり，可逆圧縮は圧縮前のデータに戻せる圧縮方法で，非可逆圧縮は圧縮前のデータに戻せない圧縮方法で，一般に非可逆圧縮のほうが容量は小さくなります．

2.4　構造化データ，非構造化データ

2.4.1　構造化データ

構造化データは，決められた構造で表現されたデータです．例として，表 1 のような売り上げデータが挙げられます．

表 1　構造化データの例 (売り上げデータ)

売り上げ NO	日付	顧客 NO	商品 NO	個数
A001	2023/09/10	C987654	SU6128	7
A002	2023/09/11	C123456	SU6128	20
A002	2023/09/11	C123456	RT7588	3

日付の列には，日付が入り，個数には，非負の整数が入るなどの決まり・制約があり，それに従って，データが入れられています．このように表現すことにより，検索・抽出（たとえば，顧客番号 C123456 の一覧）や計算・集計 (たとえば，SU6128 の個数の合計) をしたりすることが容易に行えます．構造化データは，表計算ソフトウエアでも管理できますが，より大規模な場合は，リレーショナル（関係）データベース (RDB) として管理されます．

2.4.2　非構造化データ

非構造化データは，表の形で表現することが困難データで，文書（テキスト）データ，画像データ，音声データなどがあります．分析などのデータ利用が，構造化データと比べて困難なことが多いです．しかし，さまざまな統計，数理，AI などを利用した分析・解析ツールが登場しています．

非構造化データの分析例として，次節で，KH コーダを利用したテキストの可視化を体験します．

2.5　半構造化データ

構造化データのように決まった構造ではなく，柔軟な構造を与えたものです．SVG の例で，タグ (`<circle>`）は，円を描くことを示しています．SVG では円以外にも線，矩形，文字などさまざまなタグで図形をえがくことができ，順番や個数など柔軟に表現できます．

また，web のページは，html で表現されており，右クリックして，ページのソースを表示 で HTML 表現を見ることができます．さまざなタグ，`<HTML>`や`<li>`，`<script>`などは，表示情報の種類を表していますが，テキスト同様に，順番や回数は自由です．

このように，半構造化データは，構造化データと非構造化データの中間の性質を持っています．

3　KH コーダによるテキスト分析

KH コーダは，テキスト（文書）を入力して，そのテキストを計量分析をして，関係をあらわしたネットワーク図や樹形図などで可視化します．大量のテキストデータからなんらかの知識などを取り出すことはテキストマイニングと呼ばれています．

本章での分析は，テキスト処理を体験することを目的としており，きちんとした分析 – 前処理，分析パラメータの設定，

紹介したツール以外の分析ツール – については，他の授業や参考文献による学修に譲りたいと思います．

3.1　KH コーダーでやってみること

KH コーダーは，テキスト（文書）を入力するといろいろな分析ができますが，本書では，体験として，次のことをやってみます．

- 単語の出現頻度
- 単語の前後でどのような語が出現するか？
- 共起ネットワーク図の作成
- クラスター分析により，単語をグループ分けする．

3.2　VDI の起動

KH コーダーは，大学の VDI にインストールされているので，VDI を利用します．KH コーダーは，フリーソフトウエアなので，皆さんの PC にダウンロードし，インストールできます[*1]．

本章の 1 節で示した方法で，専修大学の VDI(AVI) を起動してください．スタートメニューから Senshu-AVD-HP をクリックします．

本節では，VDI で作業をし，その結果を専修大学 OneDrive に保存します．そのため，まず，Senshu-AVD-HP にログオンしたら，VDI のウインドウで OneDrive にログオンしてください．

2023 年 9 月現在，専修大学の VDI 環境のエクスプローラでは，拡張子（ファイルの種類を示すもの）は表示されません．

[*1] KH Coder からダウンロード，インストールできます．Mac へのインストールは，UNIX システムについての十分な知識と経験が必要とのことです．

表示させるには，

エクスプローラのメニューの 表示 → 表示 → ファイル名拡張子

とします．

動画:拡張子の表示（音声なし）

3.3　分析用のファイルの作成

青空文庫から，壺井栄の「二十四の瞳」をダウンロードして分析してみたいと思います．

3.3.1　ファイルのダウンロード・解凍

(1) ブラウザで，青空文庫の壺井栄，「二十四の瞳」，図書カードのページに行きます
https://www.aozora.gr.jp/cards/001875/card57856.html

(2)「ファイルのダウンロード」中のテキストファイル (ルビあり) の 57856_ruby_63623.zip を右クリックして，名前を付けてリンク先を保存で OneDrive の適当な場所にダウンロードします．

(3) エクスプローラを起動して，57856_ruby_63623.zip を右クリックして，すべて展開をクリックして，展開をクリックします．

(4) 57856_ruby_63623 というホルダの中の nijushino_hitomi.txt が「二十四の瞳」のテキストデータになります．

動画:二十四の瞳，ダウンロード（音声なし）

3.3.2　ファイルの確認，記号の削除

ファイルを確認し，ルビや入力者注の記号を削除などをします．削除には 正規表現というツールを使いますが，ここでは正規表現の説明は省略します．正規表現を利用するには，正規表現に対応したテキストエディタを使います．専修大学の VDI には，サクラエディタがインストールされているので利用します．他の正規表現に対応したテキストエディタでもほぼ同様に利用できます．

(1) エクスプローラで，nijushino_hitomi.txt を右クリックし，プログラムから開く → サクラエディタ を選びます．

(2) 変更を加えるので，別のフォルダに保存して，それを利用します．
メニューの 名前を付けて保存 とし，保存場所を変更して，保存 をクリック

(3)「昔《むかし》」のように，ルビは「《」（全角）と「》」（全角）で囲まれています．これを削除します．
（a）メニューの 検索 → 置換
（b） 置換 のウインドウが表示されます．
（c） 正規表現 にチェックを入れます．
（d） 置換前: に，《.*?》 を，置換後は空（何も入力しない状態）にします [*2]
（e） 置換 をクリックすると 1 つ置換されます． すべてを置換 をクリックすると残りすべてが置換されます．

(4) 1 行目に戻り，同様に，「｜」（全角）を削除します．1 行目に戻り，同様に， 置換前: に ｜ を，置換後は空（何も入力しない状態）にします．

(5)「［＃」（全角）で始まり，「］」（全角）までが，入力者注を意味します．その部分を削除します．1 行目に戻ります．
置換前: に ［＃.*?］ を，置換後は空（何も入力しない状態）にします．

[*2] 参考:「《」(全角) という文字列があり，「.」（半角）は何かの文字を意味し，「*」（半角）はそれが何回出現してもよく，「》」(全角) までを置き換えることを意味します．「?」（半角）は最小一致を表し，「《」の後ろに複数の「》」があるとき，最初の「》」までを置き換えます．

(6) 記号の説明部分（この例では 4 行目から 15 行目）を削除します．

(7) 入力元の本の情報，ファイル最後の方（1938 行目付近）の「底本：」から下を削除します．

(8) 上書き保存をしておきましょう．

動画:青空文庫記号の削除（音声なし）

3.4 KH コーダでの分析

3.4.1 KH コーダーの起動

スタート → すべてのアプリ → KHcoder3

3.4.2 ファイルの読み込みとチェック

3.3.2 小節で，記号を削除したファイルを読み込みます．

(1) KH コーダーのメニューの プロジェクト → 新規

(2) 参照 のボタンを押すなどして，記号を削除したファイルを指定し，OK

(3) 言語の中の携帯素解析ソフトウエアの設定

KH コーダーでは，ChaSen と MeCab という 2 つのソフトウエアから選択できます．このテキストでは，ChaSen を使った場合で説明しています．どちらを使うかは，担当教員の指示に従ってください．

次に，前処理でテキストのチェックをします．

メニューの 前処理 → テキストのチェック

動画:KH コーダー，起動・ファイル読み込み，テキストチェック（音声なし）

3.4.3　前処理

前処理は，形態素解析という作業をし，入力した文書を単語に分けます．形態素解析は，KH コーダーのデフォルトでは茶筌 (ChaSen) または MeCab というソフトウエアが使われます．

メニューの 前処理 → 前処理の実行

3.4.4　抽出語

どのような語が多く使われているかを棒グラフで表示します．

- メニューの ツール → 抽出語 → 抽出語リスト
- 抽出語リスト のウインドウで，Excel 出力 をクリックし，そのまま OK をクリックすることで，Excel でみることができます．

出現数の多い単語を見ることで，内容をある程度予測できます．また，抽出語リスト のウインドウで，単語をクリックすれば，その単語の前後が表示されます．

動画:前処理・抽出語（音声なし）

3.4.5　共起ネットワーク

共起ネットワークは，同じ段落（デフォルトの設定）で，単語と単語が一緒に出現するかしないかをもとにネットワークを作成します．

(1) メニューの ツール → 抽出語 → 共起ネットワーク

(2) 抽出語・共起ネットワーク：オプション が表示されます.
オプションは，初期値の状態から，強い共起関係ほど濃い線に と 最小スパニング・ツーリーだけを描画 の 2 つにチェックを入れ,， OK

(3) しばらくすると, 共起ネットワーク図が表示されます.

図 5 は，例題の共起ネットワーク図です．図の見方：

- ○の大きさが語の出現頻度を表しています.
- 線で結ばれていると，語と語が共起（ともに出現）していることを示し，線が太いほど，その共起が強いことを示しています.
- サブグラフは，強く結びついている語のグループをまとめて，同じ色で表示しています.
- 注意：○の近さに意味はありません.

問題点はありますが，後ほど修正ます.
動画:共起ネットワーク（音声なし）

3.4.6 階層的クラスター分析

階層的クラスター分析は，語と語の近さから，語をクラスター（グループ）にまとめていきます.

(1) メニューの ツール → 抽出語 → 階層的クラスター分析

(2) 抽出語・クラスター分析：オプション が表示されます.
オプションは，初期値の状態のまま， OK をクリック.

(3) しばらくすると, デンドログラム（樹形図，図 7 のような図）が表示されます.

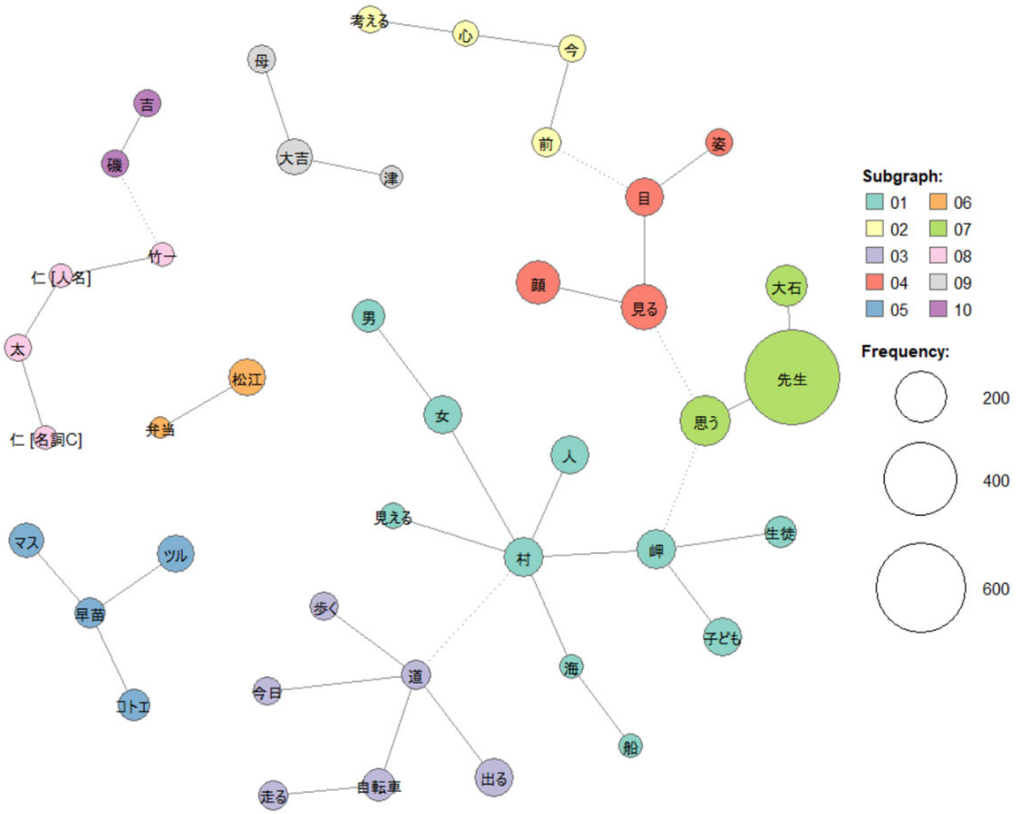

図 5　例題の共起ネットワーク図

図の見方：

- 語と語の近さで，近い語を順次併合していっています．しがって，左でつながった語がより近いことを表しています．
- 点線の値を基準で色分けをして，同じクラスターに含まれる語は同じ色になっています．
- 語をクリックすると，どのように使われいるのかが表示されます．

動画:クラスター分析（音声なし）

3.4.7　抽出語の修正

図 5 を見ると，語の誤認識がありあます．例えば，「磯吉」を「磯」と「吉」を別の単語として認識，他にも「八津」同様に誤認識しています．どのようになっているかは，図で「磯」をクリックするとみることができます．

そこで，「磯吉」などを単語として認識させるために，「強制抽出する語」に，「磯吉」などを追加します．

(1) メニューの 前処理 → 語の取捨選択
(2) オプションは，初期値の状態のまま， OK をクリック．
(3) しばらくすると, ウインドウが表示されます． 強制抽出する語の指定 に，
磯吉
八津
仁太
小ツル
マスノ
を追加します．この語の選択は，共起ネットワーク図と二十四の瞳のテキストファイルの 116 行目あたりから下を参考にしました．

(4) 語の抽出は，前処理で行うので，再度，前処理を実行します．
　　メニューの 前処理 → 前処理の実行

動画:クラスター分析（音声なし）

3.4.8　パラメータの調整

同様に，抽出語の共起ネットワークや抽出語のクラスター分析を行うことができます．

図を見ていくと，もっと多くの語の分析をしたい，もっと少ない語でシンプルな図で分析をしたいということがあります．そのときは，抽出語・共起ネットワーク：オプション や 抽出語・クラスター分析：オプション の出現数による取捨選択で，最小出現数を調整することにより，図中の語数を変更できます．語の最小出現数を大ききすると，利用される語の数が少なくなり，シンプルな図になし，逆に，語の最小出現数を小さくすると，利用される語の数が多くなり，多くの語を使った分析になります．

図 6 は，抽出語・共起ネットワーク：オプション で最小出現数で，20 にしたものです．
図 7 は，抽出語・クラスター分析：オプション で，最小出現数で 40 にしたもので，図の左右は上下でつながっています．

語の最小出現数をいろいろ調整してみましょう．

図の保存は，次のようにします．

(1) 図の右下の 保存 ボタンをクリック
(2) OneDrive の適当な場所に保存します．
(3) ファイルの種類は，選択できますが，PNG か emf が利用しやすいです．

KH コーダーの分析結果は保存されません．分析結果は，図や Excel 形式で保存します．

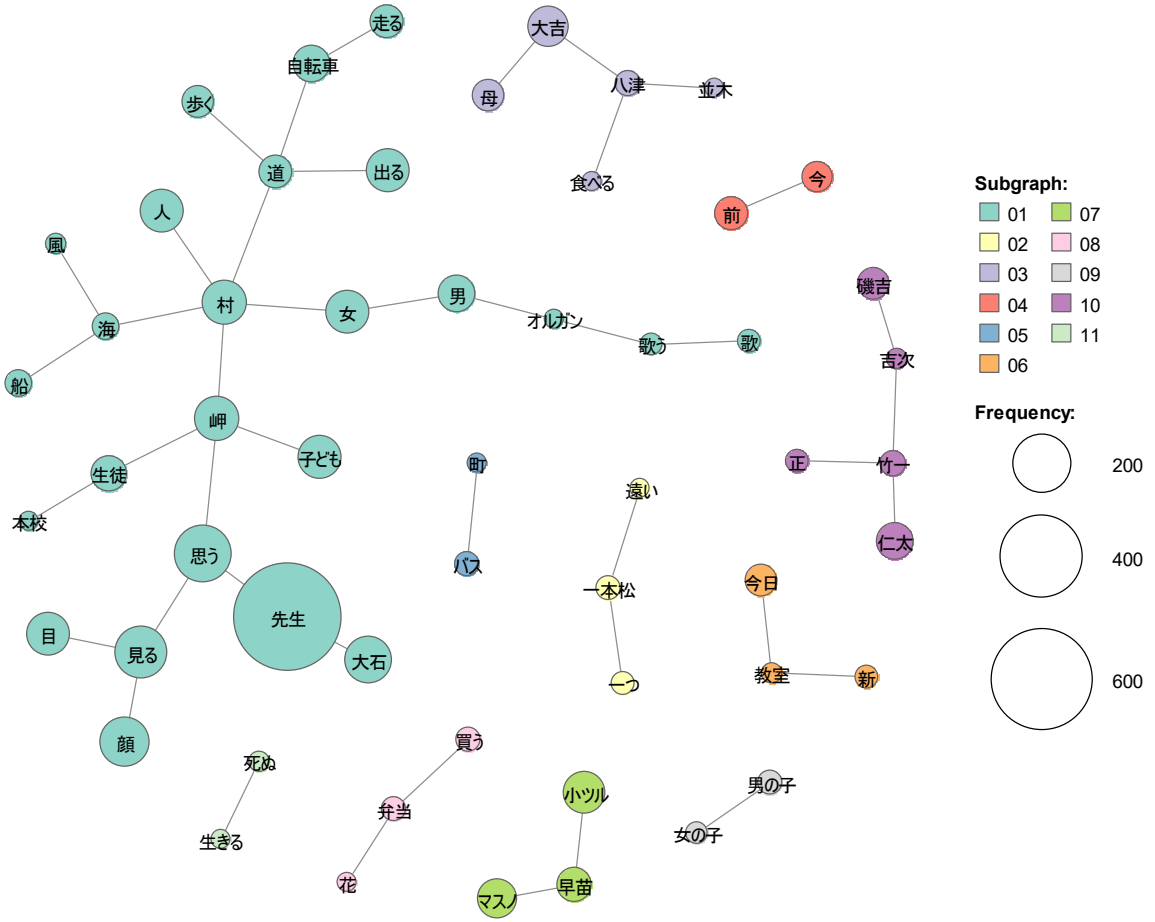

図 6　抽出語の修正し，最小出現数を 20 にした共起ネットワーク図

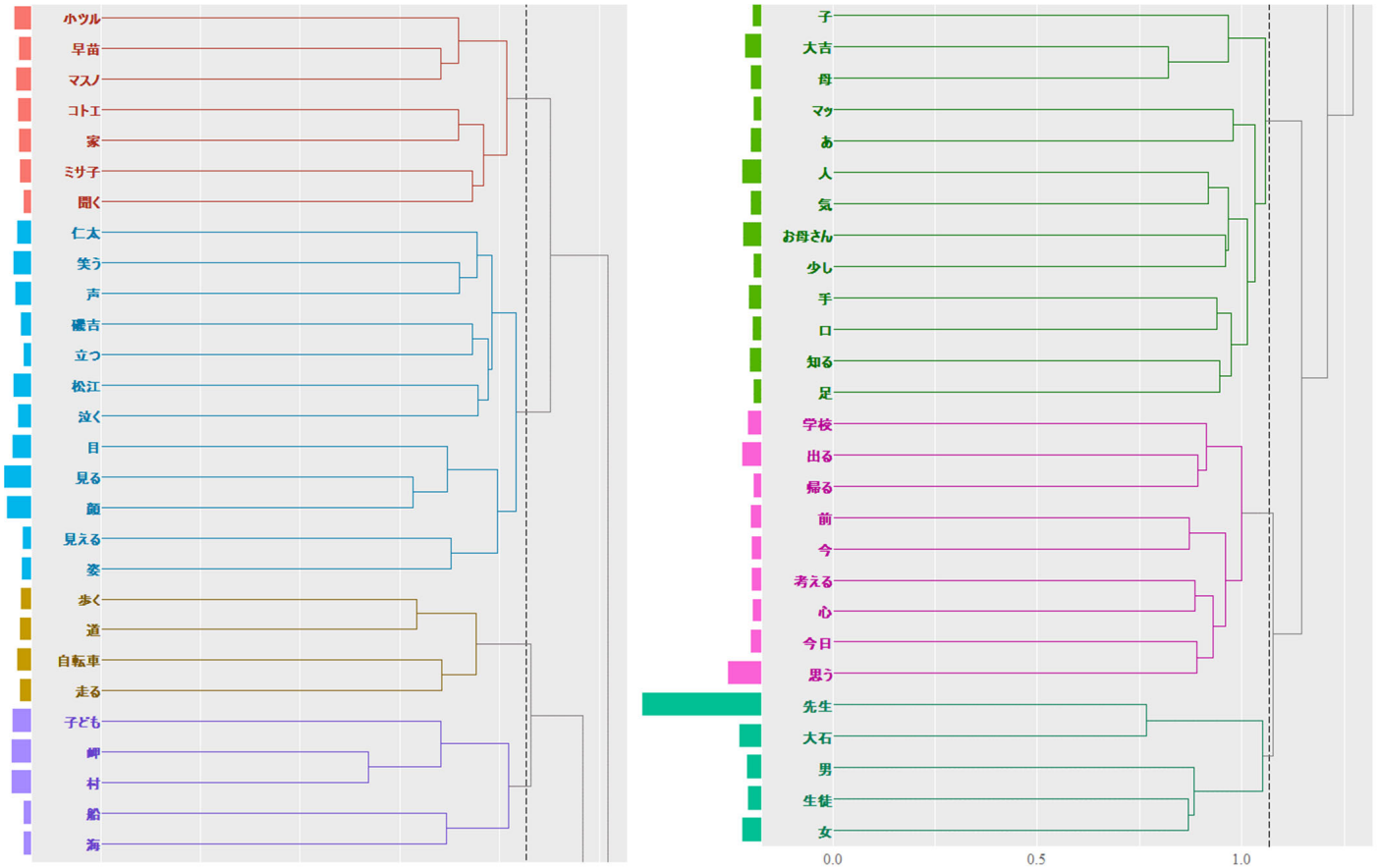

図 7　抽出語の修正し，最小出現数を 40 にした階層クラスタ図

3.4.9　図の解釈

共起ネットワーク図 (図 6 の例):

- 01 のサブグラフで，村の様子が出現しています．そのなかには，「先生」を中心とする部分と「村」，「岬」，「道」が「村」を中心に結びつき，「村」，「岬」，「道」もまた，中心になっています．
- 名前のサブグラフが 3 つ（「大吉」，「竹一」「早苗」が中心）出現しています．それぞれが何らかの関係があることが見えてきます．
- 「死ぬ」と「生きる」がつながっており，生死に関わる話が出現していることを示しています．
- 「男の子」と「女の子」のサブグラフ，「女」と「男」が線で結ばれています．性別を意識した話であるか，性別を意識せざるおえない時代背景の話ではないかと考えられます．

階層クラスタ図 (図 7 の例):

- 左上の赤の部分で，「早苗」と「マスノ」がまず結びつき，そこに「小ツル」が結びついていることが分かります．
- 赤の部分のクラスターは，「早苗」，「マスノ」，「小ツル」のクラスターと「コトノ」，「ミサ子」のクラスターからなることが分かります．
- 左の水色のクラスターでは，「仁太」，「磯吉」，「松江」の小クラスターがあり，それがまとまって中クラスターになっている．それぞれの小クラスター内で結びついている語で，その人の様子が想像できる．また，「目」，「見る」，...，「姿」の中クラスターがあり，それが，最初の中クラスターに結びついている．
- 他にもいろいろ読み取れると思います．

3.4.10　共起の強さ (類似係数) の計測方法 – Jaccard 係数 –

語と語が共起する強さ (類似係数) は，１つの段落（この例）を単位とした Jaccard 係数を使っています（デフォルト）．$|A \cap B|$ を語 A,B がともに出現する段落数，$|A \cup B|$ を語 A,B が少なくてもどちらかが出現する段落数とします．語 A と B の Jaccard 係数は，次の式で計算されます．

$$\text{語 } A \text{ と } B \text{ の Jaccard 係数} = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{ともに出現数}}{\text{どちらかの語の出現数}}$$

語 A が 20 個の段落に出現し，語 B が 60 個の段落に出現し，語 A と B がともに出現する段落数を 15 とします．（語 A と B の少なくてももどちらかが出現する段落は，$20 + 60 - 15 = 65$ になります）．このときの Jaccard 係数は，

$$\text{語 } A \text{ と } B \text{ の Jaccard 係数} = \frac{15}{20 + 60 - 15} = \frac{15}{65} = 0.25$$

となります．

3.5　分析の練習

なにかの文書の分析をしてみましょう．例えば，青空文庫から，ダウンロードして，「二十四の瞳」と同様の手順で分析できます．ただし，青空文庫は基本的に「作者の死後 50 年を経るなどして著作権の消滅した作品」を扱ってます．

3.6　図などの利用

図やテキストファイルは，OneDrive に保存しました．OneDrive に保存してあれば，自分の PC の OneDrive, 専修大学の VDI での両方で利用できます．専修大学 VDI で保存したファイルを，専修大学 VDI（リモートデスクトップ）を切断後，自分の PC の OneDrive からファイルを表示してみましょう．

4　エクスポート・削除・終了

KH コーダーでは，プロジェクトを保存できます．分析を続けたいときは，エクスポート機能を使って，プロジェクトを保存します．

メニューの[プロジェクト] → [エクスポート] → [KH コーダー形式 (インポート可)]

を選び，OneDrive に保存します．読み込むときは，

(1) メニューの[プロジェクト] → [インポート] → [KH コーダー形式]
とし，読む込むファイルを選択します．
(2) 前処理が行われます．
(3) [プロジェクトマネージャー]がされるので，自分の ID を含むディレクトリのファイルを選び，[開く]をクリック

とします．

終了する前に，現在のプロジェクトを削除する必要があります（2023/10/02 現在，削除しないと他の利用者 が参照できることがあります）．

(1) メニューの[プロジェクト] → [開く]
(2) 自分のファイル：ディレクトリ（フォルダー）に自分の ID があるファイルを選択
(3) [削除]をクリック

終了方法：

メニューの[プロジェクト] → [終了]